Sādhanā (2022)47:123
https://doi.org/10.1007/s12046-022-01878-3

Kinnauri-Pahari (version_0.1): parallel, monolingual dataset and word-embeddings

SHEFALI SAXENA*, SHWETA CHAUHAN and PHILEMON DANIEL

Electronics and Communication Department, National Institute of Technology, Hamirpur, H. P., India
e-mail: shefali@nith.ac.in; shweta@nith.ac.in; phildani7@nith.ac.in

MS received 2 March 2022; revised 24 March 2022; accepted 30 March 2022

Abstract. The recent United Nations Educational, Scientific and Cultural Organization (UNESCO) survey states that India has 197 endangered languages. Himachal Pradesh, a state in India, has topped the list with seven definitely endangered languages, and Kinnauri-Pahari being the one. Due to the lack of availability of digitized resources, the corpus compilation is a bit difficult. This paper presents and releases the Kinnauri-Pahari (ISO-639-3:kjo) dataset, consisting of the 43,362 Monolingual and 20,307 Parallel sentences in version_0.1. The dataset was tested on the Statistical, and Neural Machine Translation and their results were evaluated using different evaluation metrics. The corpus is freely available for non-commercial usage and research (https://github.com/phildani7/dlnith/tree/master/Kinnauri-Pahari).

Keywords. Low Resource Language; Machine Translation; Evaluation.

1. Introduction

Due to the pervasive digital presence and use in Natural Language Processing (NLP) applications, compiling high resource languages data is no longer difficult. Standard corpora have been created for these high-resource languages, whereas under-resourced languages have yet to establish a digital presence. Low resource languages have insignificant or no datasets to use supervised learning techniques.

The United Nations Educational, Scientific and Cultural Organization (UNESCO) lists 2464 endangered languages throughout the world [1]. India outshines the list with 197 endangered languages, accompanied by the United States (191) and Brazil (190). India is a vastly diverse country in religion, culture, language, and art, with 81 vulnerable languages, 63 definitely endangered, 6 severely endangered, and 42 critically endangered. Children's no longer learn the language as their mother tongue in endangered languages.

Kinnaur, Himachal Pradesh's third-largest district by geographical area [2], has a population of 84,121 people [3], is designated as a Scheduled Area and is home to nine native languages [4, 5] as depicted in figure 1. Along with the north of the Satluj River to Morang and villages from Badhal to Sangla, roughly 45,000 people speaks Kinnauri. A thousand people in Chitkul and Rakchham villages of Sanga valley near Baspa river speaks Chitkul Kinnaur. Almost 2500 people speak Sumcho in Pootehsil, Kanam, Spilo, Labrang, Taling, Rushkaling, and Shyaso villages. The Poo division of upper Kinnaur speaks Bhoti Kinnauri (Tibetan dialect); it is spoken in the Nakoand Hangrang villages by roughly 7000 people. Chhoyuli is a Tibetic language spoken in Kunnu, Nesang and Charan villages in upper Kinnaur's Poo district. It has a population of approximately 700 people. Sunnam village in the upper Kinnaur of Poo division speaks the Sunnami language. It has a population of around 700 people [6].

Kinnauri-Pahari (ISO-639-3:kjo), also known as Oras Boli is written in the Takri script, and is one of seven endangered languages found in Himachal Pradesh, India. Kinnauri-Pahari is an Indo-Aryan language variation spoken mainly by the Scheduled Caste community in the Kinnaur tehsils of Sangla, Moorang, Nichar, and Kalpa villages of Kinnaur districts of Himachal Pradesh. It is noted for its linguistic richness, with linguistic variations from the Sino-Tibetan (Tibeto-Burman) and Indo-European language families coexisting [7, 8]. However, these villages are home to fewer than 10,000 people [6], with just a tiny percentage of proficient speakers as per the Census 2011 [9].

Kinnauri-Pahari is not an official language in offices, tourism, or education, therefore, deteriorating every year. As a result, there is an immediate need to improve language usage in everyday situations. To achieve the purpose, the digitalization of the language and the development of machine translation models, digital dictionaries, and other tools are critical. The challenge rises as Kinnauri-Pahari becomes more morphologically rich and lacks well-defined

*For correspondence
Published online: 27 June 2022

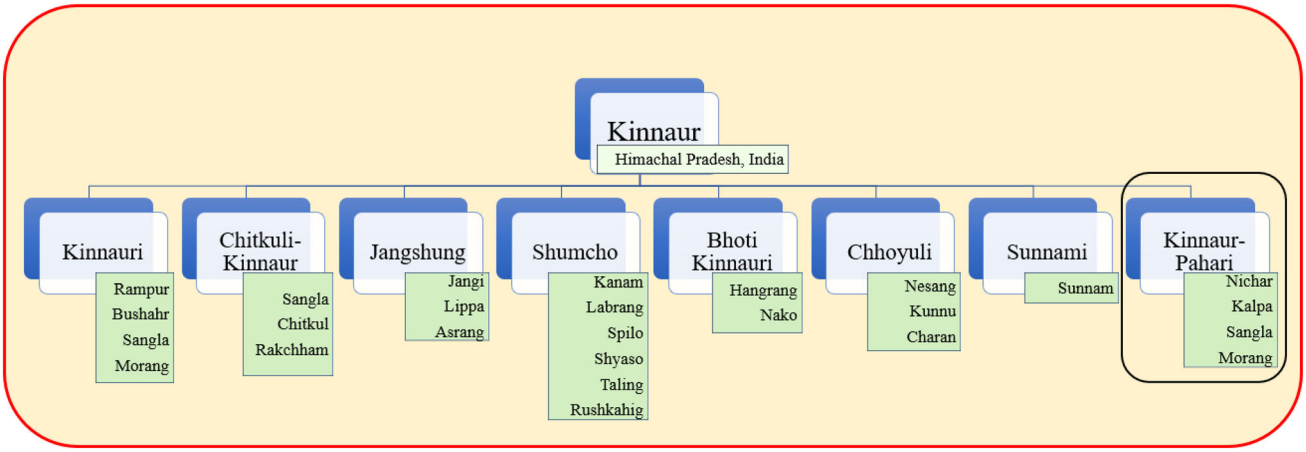

Figure 1. Illustrates the linguistic varieties and geographical distribution of Kinnaur District.

linguistic rules. Compiling a corpus is difficult because dialect differs slightly from region to region.

The paper presents the Kinnauri-Pahari monolingual and Hindi-Kinnauri-Pahari parallel datasets. The Kinnauri-Pahari dataset is manually assembled and structured. As a result, the monolingual dataset has 43,362 sentences, while the parallel dataset has 20,307 sentences. FastText [10] is also used to share pre-trained Kinnauri-Pahari word embeddings. For SMT and NMT translations, we have provided the BLEU and METEOR evaluation scores.

2. Data collection

The Indian dataset has been collected by using data extraction technique like, the IITB corpus [11] compiled the English-Hindi parallel corpus from a variety of existing sources and classified them in the following categories: OPUS, HindEn, Judicial domain corpus-I, Judicial domain corpus-II, Hindi-English Linked Wordnets, TED talks Indic Multi-parallel corpus, Wiki Headlines, Gyaan-Nidhi Corpus as well as corpora developed at the Center for Indian Language Technology, IITB. The IndicNLP [12] introduce monolingual corpus for 11 major Indian languages. The corpora are collected by crawling news articles, magazines and blogposts. Data is sourced from popular Indian language news websites. Most sources through online newspaper directories (e.g., w3newspaper) and automated web searches using hand-picked terms in various languages. After content extraction, applied filters on content length, script, etc., to select good quality articles. [13] offer the collection of parallel corpora for Indian languages. They mined the parallel sentences from the web by combining many corpora, tools, and methods such as web-crawled monolingual corpora, document OCR for extracting sentences from scanned documents, and approximate nearest neighbor search for searching in a large collection of sentences.Since this Kinnauri-Pahari corpus is a endangered low-resource language with no digitalization and since there are fewer people who know both Hindi and Kinnauri-Pahari, preparing corpus and compiling dataset is a challenging task. As a result, data extraction from web resources is not available. The dataset was manually prepared by translators, and also the word embeddings has been provided. The corpus statistics for the translation task for the language are described in table 1.

2.1 *Creation of monolingual dataset*

The sentences for monolingual corpus compilation are manually corrected, it is time consuming and tedious task. The monolingual dataset contains various short and long stories and daily life conversations.

2.2 *Creation of parallel Hindi-Kinnauri-Pahari dataset*

The Parallel Hindi-Kinnauri-Pahari dataset is shown in table 1. The parallel corpus has been created by distributing different hindi stories and everyday topics to the writers. Both Hindi and Kinnauri-Pahari are written from scratch.

Table 1. Kinnauri-Pahari corpus statistics.

Dataset	Sentence	Vocabulary	Train	Test
Monolingual				
Kinnauri-Pahari	43,362	50,880	40,360	3000
Parallel				
Hindi	20,307	12,377	19,804	500
Kinnauri-pahari	20,307	27,748	19,804	500

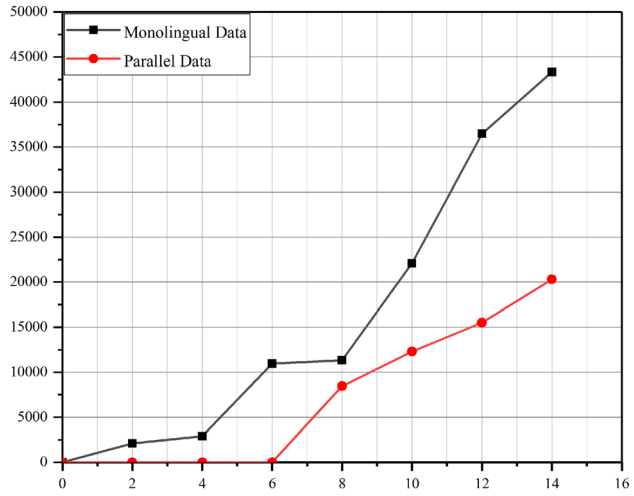

Figure 2. The total volume of monolingual and parallel data collected as a function of elapsed months for Kinnauri-Pahari.

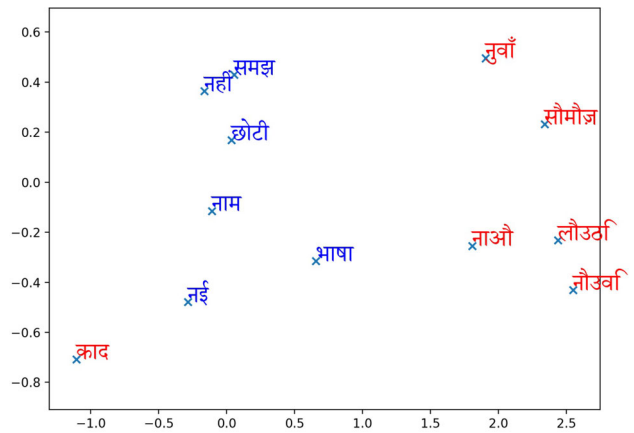

Figure 3. The shared embedding space of Hindi and Kinnauri-Pahari.

The main categories of dataset include data from the following domains: Hospital, Market, Defense, Food, Parties, Media, School, Law, Marriage, Music, Education, Technology, Sports, Culture, History, Dance, Religion, Stories, etc.

The amount of data collected as a function of the amount of time from posting the task is depicted in figure 2.

3. Word embedding

Word embedding is a vector representation of each word that can capture its context, syntactic and semantic similarity, analogies, and relationships with other words. Fast Text [10] is a method based on the skip-gram and CBOW models, in which each word is represented as a bag of n-gram characters. FastText was used to tokenize the corpus since it has a better word relationship. For crosslingual common embedding space, vecmap [14] is used. The word embedding representation in two-dimensional space is shown in figure 3. It depicts the Hindi and Kinnauri-Pahari shared embedding space. The word embedding training settings are set to 10 epochs, 0.05 learning rate, and 300 word embedding dimension. Using the closest neighbour approach, several Kinnauri-Pahari words in the mapped region are displayed. Principle Component Analysis (PCA) is used to decrease the dimensionality of certain random words from various domains. As seen in figure 3, several Hindi-Kinnauri-Pahari terms, such as "नाम" and "नाओ", are highly similar to one another and may be simply translated from one language to another. The table 2 shows the prediction of best five nearest neighbor for Kinnauri-pahari words along with their similarity scores. The table 3 predicts the best five values for Hindi source words to the Kinnauri-Pahari target space.

4. Machine translation

The section shows the experimental setup created for two Machine translation models used for corpus analysis.

4.1 *SMT*

We used the Phrase-Based SMT (PBSMT) system for training SMT systems using monolingual corpora [15]. We used the default settings of Moses for the experiments.

Table 2. The five Nearest Neighbor of Kinnauri-Pahari word-word translation.

Source Word	थीह		होतेसै		कियै		मैईयै	
Sr. No	Word	Value	Word	Value	Word	Value	Word	Value
1.	थीह	1.0000	हौतेसै	1.0000	कियै	1.0000	मैईयै	1.0000
2.	थी,	0.8727	हौतेसलै	0.9676	कियैन	0.9456	तैईयै	0.9580
3.	थीह,	0.8529	हौतेसा	0.9674	"कियै	0.9210	मैई	0.9321
4.	थीह।	0.8484	हौतेसाई	0.9470	कियैए	0.9090	करटकै	0.8940
5.	दाखनी	0.8446	पिंगलकै	0.9294	कियू	0.8996	हेबरै	0.8825

Table 3. The five Nearest Neighbor of Hindi to Kinnauri-Pahari word-word translation.

Source Word	तुम		दुखी		भाषा		ठीक	
Sr. No	Word	Value	Word	Value	Word	Value	Word	Value
1.	दितांह	0.6272	ईज़ित	0.6564	औबौल	0.5676	ज़ुनिये	0.4366
2.	तांह	0.6074	तरसन	0.6552	भौली	0.5589	ऐनोरिये	0.4357
3.	हुतांह	0.5736	सोंसार	0.6471	खेलना	0.5557	ताम्बिये	0.4348
4.	कितांह	0.5674	शिख्यानो	0.6468	रौक	0.5552	निथियोह,	0.4319
5.	दिसुहं	0.5626	बुरिङ	0.6442 -	कौरना,	0.5484	तेनोरिये	0.4315

Table 4. Machine translated result of Hindi Kinnauri-Pahari for BLEU and METEOR evaluation metrics.

	Hindi → Kinnauri-Pahari		Kinnauri-Pahari → Hindi	
	BLEU	METEOR	BLEU	METEOR
NMT	9.3	3.5	12.4	3.6
SMT	11.4	4.6	16.2	4.9

4.2 *NMT*

The NMT model was trained using the shared encoder with backtranslation [17] that consists of a shared encoder with two different decoders. We have used two layered encoder decoders with [19], 600 hidden units each, maximum sentence length of 50, batch size of 50, and Adam optimizer is used with a learning rate of 0.0002.

4.3 *Evaluation*

We evaluated machine-translated output using BLEU [16] and METEOR [18] evaluation metrics. Table 4 shows the results of our experiments.

5. Conclusion

In this paper, the corpus of low-resource Himachali language, i.e., Kinnauri-Pahari version_0.1 language, has been released. The paper presents the statistical and neural MT results for Kinnauri-Pahari version_0.1. The corpus is publicly available under a creative commons license.

In future, the plan is to enhance the Kinnauri-Pahari corpus and also provide other low resources, endangered Himachali languages datasets. We anticipate that making this corpus available would help speed up NLP research in Indian languages by allowing the community to provide more resources and solutions for various NLP jobs and offer up new NLP questions.

We hope that this research will be a first step towards documenting, preserving and developing the Kinnauri Pahari language for future generations.

Acknowledgements

The Project, Documentation of Kinnauri Pahari dialect and development of artificial intelligence based translation of Kinnauri Pahari, is an year project and is funded by the Tribal Development Department, Government of Himachal Pradesh.

References

[1] Unesco.org 2017 http://www.unesco.org/languagesatlas/ [Online; accessed 04-July-2017]
[2] Census of India 2011 Kinnauri District:Census 2011 Data, Accessed on 27th July, 2015 https://www.census2011.co.in/census/district/240-kinnaur.html
[3] Census of India 2011 Himachal Pradesh Population Census Data 2011 https://www.census2011.co.in/census/state/himachal+pradesh.html
[4] Lewis M P, Simons G F and Fenning C D 2015 (eds) Ethnologue: languages of the World, Eighteenth edition. Dallas, Texas: SIL International. Online Version: https://www.ethnologue.com
[5] Chamberlain B, Wendy C and Emma P 1998 A Sociolinguistic Survey of Kinnauri
[6] https://en.wikipedia.org/wiki/Kinnauri_language
[7] Huber C & der Wissenschaften A 2007 Researching local languages in Kinnaur, Verlag der Österreichischen Akademie der Wissenschaften,755: 249-266
[8] Kumar M A and Bezily M S 2015 Kinnauri Pahari
[9] https://censusindia.gov.in/2011census/C-16_Town.html
[10] Joulin A, Grave E, Bojanowski P and Mikolov T 2016 Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759
[11] Kunchukuttan A, Mehta P and Bhattacharyya P 2017 The IIT Bombay english-hindi parallel corpus. arXiv preprint arXiv:1710.02855
[12] Kakwani D, Kunchukuttan A, Golla S, Gokul N C, Bhattacharyya A, Khapra M M and Kumar P 2020 IndicNLPSuite: monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In: *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 4948–4961)
[13] Ramesh G, Doddapaneni S, Bheemaraj A, Jobanputra M, AK R Sharma A, and Khapra M S 2022 Samanantar: the

largest publicly available parallel corpora collection for 11 indic languages. *Trans. Assoc. Comput. Linguist.* 10: 145–162

[14] Artetxe M, Labaka G and Agirre E 2018 A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings arXiv preprint arXiv: 1805.06297

[15] Mikel Artetxe G L and Agirre E 2018 Unsupervised Statistical Machine Translation. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3632–3642

[16] Papineni K, Roukos S, Ward T and Zhu W J 2002 BLEU: a method for automatic evaluation of machine translation. In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (pp. 311–318). Association for Computational Linguistics

[17] Artetxe M, Labaka G, Agirre E and Cho K 2017 Unsupervised neural machine translation. arXiv preprint arXiv:1710.11041

[18] Banerjee S and Lavie A 2005 METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: *Proceedings of the acl Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization* (pp. 65–72)

[19] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A Kasier L and Polosukhin I 2017 Attention is all you need. In: *Advances in Neural Information Processing Systems* (pp. 5998–6008)